DUXBURY

SEVENTH EDITION

Probability and Statistics for Engineering and the Sciences

JAY L. DEVORE

California Polytechnic State University, San Luis Obispo

a great deal of uncertainty concerning the value of what we are estimating. Figure 7.1 shows 95% confidence intervals for true average breaking strengths of two different brands of paper towels. One of these intervals suggests precise knowledge about μ, whereas the other suggests a very wide range of plausible values.

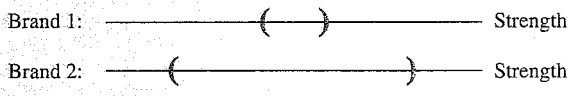

Figure 7.1 Confidence intervals indicating precise (brand 1) and imprecise (brand 2) information about μ

7.1 Basic Properties of Confidence Intervals

The basic concepts and properties of confidence intervals (CIs) are most easily introduced by first focusing on a simple, albeit somewhat unrealistic, problem situation. Suppose that the parameter of interest is a population mean μ and that

1. The population distribution is normal
2. The value of the population standard deviation σ is known

Normality of the population distribution is often a reasonable assumption. However, if the value of μ is unknown, it is implausible that the value of σ would be available (knowledge of a population's center typically precedes information concerning spread). In later sections, we will develop methods based on less restrictive assumptions.

Example 7.1 Industrial engineers who specialize in ergonomics are concerned with designing workspace and devices operated by workers so as to achieve high productivity and comfort. The article "Studies on Ergonomically Designed Alphanumeric Keyboards" (*Human Factors*, 1985: 175–187) reports on a study of preferred height for an experimental keyboard with large forearm–wrist support. A sample of $n = 31$ trained typists was selected, and the preferred keyboard height was determined for each typist. The resulting sample average preferred height was $\bar{x} = 80.0$ cm. Assuming that the preferred height is normally distributed with $\sigma = 2.0$ cm (a value suggested by data in the article), obtain a CI for μ, the true average preferred height for the population of all experienced typists. ■

The actual sample observations $x_1, x_2, \ldots, x_n$ are assumed to be the result of a random sample $X_1, \ldots, X_n$ from a normal distribution with mean value μ and standard deviation σ. The results of Chapter 5 then imply that irrespective of the sample size n, the sample mean $\bar{X}$ is normally distributed with expected value μ and standard deviation $\sigma/\sqrt{n}$. Standardizing $\bar{X}$ by first subtracting its expected value and then dividing by its standard deviation yields the standard normal variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{7.1}$$

Because the area under the standard normal curve between −1.96 and 1.96 is .95,

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = .95 \tag{7.2}$$

Now let's manipulate the inequalities inside the parentheses in (7.2) so that they appear in the equivalent form $l < \mu < u$, where the endpoints l and u involve $\bar{X}$ and $\sigma/\sqrt{n}$. This is achieved through the following sequence of operations, each yielding inequalities equivalent to the original ones.

1. Multiply through by $\sigma/\sqrt{n}$:

$$-1.96 \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

2. Subtract $\bar{X}$ from each term:

$$-\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

3. Multiply through by −1 to eliminate the minus sign in front of μ (which reverses the direction of each inequality):

$$\bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} > \mu > \bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

that is,

$$\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

The equivalence of each set of inequalities to the original set implies that

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95 \tag{7.3}$$

The event inside the parentheses in (7.3) has a somewhat unfamiliar appearance; previously, the random quantity has appeared in the middle with constants on both ends, as in $a \le Y \le b$. In (7.3) the random quantity appears on the two ends, whereas the unknown constant μ appears in the middle. To interpret (7.3), think of a **random interval** having left endpoint $\bar{X} - 1.96 \cdot \sigma/\sqrt{n}$ and right endpoint $\bar{X} + 1.96 \cdot \sigma/\sqrt{n}$. In interval notation, this becomes

$$\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \tag{7.4}$$

The interval (7.4) is random because the two endpoints of the interval involve a random variable. It is centered at the sample mean $\bar{X}$ and extends $1.96\sigma/\sqrt{n}$ to each side of $\bar{X}$. Thus the interval's width is $2 \cdot (1.96) \cdot \sigma/\sqrt{n}$, which is not random; only the location of the interval (its midpoint $\bar{X}$) is random (Figure 7.2). Now (7.3) can be paraphrased as "*the probability is .95 that the random interval* (7.4) *includes or covers the true value of* μ." Before any experiment is performed and any data is gathered, it is quite likely that μ will lie inside the interval (7.4).

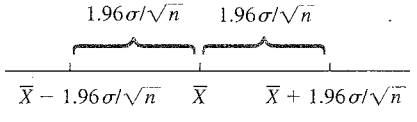

Figure 7.2 The random interval (7.4) centered at $\bar{X}$

DEFINITION

If after observing $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, we compute the observed sample mean $\bar{x}$ and then substitute $\bar{x}$ into (7.4) in place of $\bar{X}$, the resulting fixed interval is called a **95% confidence interval for μ.** This CI can be expressed either as

$$\left(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \quad \text{is a 95\% CI for } \mu$$

or as

$$\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \quad \text{with 95\% confidence}$$

A concise expression for the interval is $\bar{x} \pm 1.96 \cdot \sigma/\sqrt{n}$, where $-$ gives the left endpoint (lower limit) and $+$ gives the right endpoint (upper limit).

Example 7.2 (Example 7.1 continued)

The quantities needed for computation of the 95% CI for true average preferred height are $\sigma = 2.0$, $n = 31$, and $\bar{x} = 80.0$. The resulting interval is

$$\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} = 80.0 \pm (1.96)\frac{2.0}{\sqrt{31}} = 80.0 \pm .7 = (79.3, 80.7)$$

That is, we can be highly confident, at the 95% confidence level, that $79.3 < \mu < 80.7$. This interval is relatively narrow, indicating that μ has been rather precisely estimated. ■

Interpreting a Confidence Interval

The confidence level 95% for the interval just defined was inherited from the probability .95 for the random interval (7.4). Intervals having other levels of confidence will be introduced shortly. For now, though, consider how 95% confidence can be interpreted.

Because we started with an event whose probability was .95—that the random interval (7.4) would capture the true value of μ—and then used the data in Example 7.1 to compute the CI (79.3, 80.7), it is tempting to conclude that μ is within this fixed interval with probability .95. But by substituting $\bar{x} = 80.0$ for $\bar{X}$, all randomness disappears; the interval (79.3, 80.7) is not a random interval, and μ is a constant (unfortunately unknown to us). It is therefore *incorrect* to write the statement $P(\mu \text{ lies in } (79.3, 80.7)) = .95$.

A correct interpretation of "95% confidence" relies on the long-run relative frequency interpretation of probability: To say that an event A has probability .95 is to say that if the experiment on which A is defined is performed over and over again, in the long run A will occur 95% of the time. Suppose we obtain another sample of typists' preferred heights and compute another 95% interval. Then we consider repeating this for a third sample, a fourth sample, a fifth sample, and so on. Let A be the event that $\bar{X} - 1.96 \cdot \sigma/\sqrt{n} < \mu < \bar{X} + 1.96 \cdot \sigma/\sqrt{n}$. Since $P(A) = .95$, in the long run 95% of our computed CIs will contain μ. This is illustrated in Figure 7.3, where the vertical line cuts the measurement axis at the true (but unknown) value of μ. Notice that of the 11 intervals pictured, only intervals 3 and 11 fail to contain μ. In the long run, only 5% of the intervals so constructed would fail to contain μ.

According to this interpretation, the confidence level 95% is not so much a statement about any particular interval such as (79.3, 80.7). Instead it pertains to what would happen if a very large number of like intervals were to be constructed using the

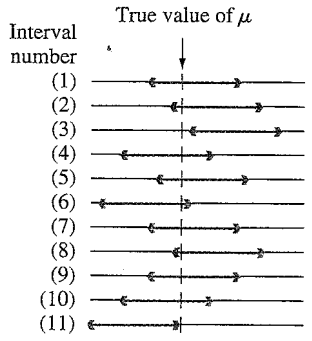

Figure 7.3 Repeated construction of 95% CIs

same CI formula. Although this may seem unsatisfactory, the root of the difficulty lies with our interpretation of probability—it applies to a long sequence of replications of an experiment rather than just a single replication. There is another approach to the construction and interpretation of CIs that uses the notion of subjective probability and Bayes' theorem, but the technical details are beyond the scope of this text; the book by DeGroot, et al. (see the Chapter 6 bibliography) is a good source. The interval presented here (as well as each interval presented subsequently) is called a "classical" CI because its interpretation rests on the classical notion of probability (though the main ideas were developed as recently as the 1930s).

Other Levels of Confidence

The confidence level of 95% was inherited from the probability .95 for the initial inequalities in (7.2). If a confidence level of 99% is desired, the initial probability of .95 must be replaced by .99, which necessitates changing the z critical value from 1.96 to 2.58. A 99% CI then results from using 2.58 in place of 1.96 in the formula for the 95% CI.

This suggests that any desired level of confidence can be achieved by replacing 1.96 or 2.58 with the appropriate standard normal critical value. As Figure 7.4 shows, a probability of $1 - \alpha$ is achieved by using $z_{\alpha/2}$ in place of 1.96.

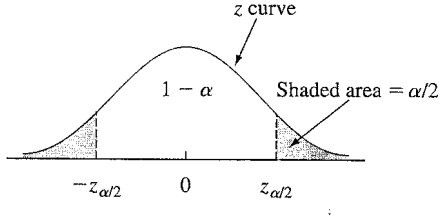

Figure 7.4 $P(-z_{\alpha/2} \le Z < z_{\alpha/2}) = 1 - \alpha$

DEFINITION

A **100(1 − α)% confidence interval** for the mean μ of a normal population when the value of σ is known is given by

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \tag{7.5}$$

or, equivalently, by $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$.

8.1 Hypotheses and Test Procedures

A **statistical hypothesis,** or just *hypothesis,* is a claim or assertion either about the value of a single parameter (population characteristic or characteristic of a probability distribution), about the values of several parameters, or about the form of an entire probability distribution. One example of a hypothesis is the claim $\mu = .75$, where μ is the true average inside diameter of a certain type of PVC pipe. Another example is the statement $p < .10$, where p is the proportion of defective circuit boards among all circuit boards produced by a certain manufacturer. If μ_1 and μ_2 denote the true average breaking strengths of two different types of twine, one hypothesis is the assertion that $\mu_1 - \mu_2 = 0$, and another is the statement $\mu_1 - \mu_2 > 5$. Yet another example of a hypothesis is the assertion that the stopping distance under particular conditions has a normal distribution. Hypotheses of this latter sort will be considered in Chapter 14. In this and the next several chapters, we concentrate on hypotheses about parameters.

In any hypothesis-testing problem, there are two contradictory hypotheses under consideration. One hypothesis might be the claim $\mu = .75$ and the other $\mu \neq .75$, or the two contradictory statements might be $p \geq .10$ and $p < .10$. The objective is to decide, based on sample information, which of the two hypotheses is correct. There is a familiar analogy to this in a criminal trial. One claim is the assertion that the accused individual is innocent. In the U.S. judicial system, this is the claim that is initially believed to be true. Only in the face of strong evidence to the contrary should the jury reject this claim in favor of the alternative assertion that the accused is guilty. In this sense, the claim of innocence is the favored or protected hypothesis, and the burden of proof is placed on those who believe in the alternative claim.

Similarly, in testing statistical hypotheses, the problem will be formulated so that one of the claims is initially favored. This initially favored claim will not be rejected in favor of the alternative claim unless sample evidence contradicts it and provides strong support for the alternative assertion.

DEFINITION

The **null hypothesis,** denoted by H_0, is the claim that is initially assumed to be true (the "prior belief" claim). The **alternative hypothesis,** denoted by H_a, is the assertion that is contradictory to H_0.

The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that H_0 is false. If the sample does not strongly contradict H_0, we will continue to believe in the truth of the null hypothesis. The two possible conclusions from a hypothesis-testing analysis are then *reject* H_0 or *fail to reject* H_0.

A **test of hypotheses** is a method for using sample data to decide whether the null hypothesis should be rejected. Thus we might test H_0: $\mu = .75$ against the alternative H_a: $\mu \neq .75$. Only if sample data strongly suggests that μ is something other than .75 should the null hypothesis be rejected. In the absence of such evidence, H_0 should not be rejected, since it is still quite plausible.

Sometimes an investigator does not want to accept a particular assertion unless and until data can provide strong support for the assertion. As an example, suppose a company is considering putting a new type of coating on bearings that it produces. The true average wear life with the current coating is known to be

1000 hours. With μ denoting the true average life for the new coating, the company would not want to make a change unless evidence strongly suggested that μ exceeds 1000. An appropriate problem formulation would involve testing H_0: $\mu = 1000$ against H_a: $\mu > 1000$. The conclusion that a change is justified is identified with H_a, and it would take conclusive evidence to justify rejecting H_0 and switching to the new coating.

Scientific research often involves trying to decide whether a current theory should be replaced by a more plausible and satisfactory explanation of the phenomenon under investigation. A conservative approach is to identify the current theory with H_0 and the researcher's alternative explanation with H_a. Rejection of the current theory will then occur only when evidence is much more consistent with the new theory. In many situations, H_a is referred to as the "researcher's hypothesis," since it is the claim that the researcher would really like to validate. The word *null* means "of no value, effect, or consequence," which suggests that H_0 should be identified with the hypothesis of no change (from current opinion), no difference, no improvement, and so on. Suppose, for example, that 10% of all circuit boards produced by a certain manufacturer during a recent period were defective. An engineer has suggested a change in the production process in the belief that it will result in a reduced defective rate. Let p denote the true proportion of defective boards resulting from the changed process. Then the research hypothesis, on which the burden of proof is placed, is the assertion that $p < .10$. Thus the alternative hypothesis is H_a: $p < .10$.

In our treatment of hypothesis testing, H_0 will always be stated as an equality claim. If θ denotes the parameter of interest, the null hypothesis will have the form H_0: $\theta = \theta_0$, where θ_0 is a specified number called the *null value* of the parameter (value claimed for θ by the null hypothesis). As an example, consider the circuit board situation just discussed. The suggested alternative hypothesis was H_a: $p < .10$, the claim that the defective rate is reduced by the process modification. A natural choice of H_0 in this situation is the claim that $p \geq .10$, according to which the new process is either no better *or* worse than the one currently used. We will instead consider H_0: $p = .10$ versus H_a: $p < .10$. The rationale for using this simplified null hypothesis is that any reasonable decision procedure for deciding between H_0: $p = .10$ and H_a: $p < .10$ will also be reasonable for deciding between the claim that $p \geq .10$ and H_a. The use of a simplified H_0 is preferred because it has certain technical benefits, which will be apparent shortly.

The alternative to the null hypothesis H_0: $\theta = \theta_0$ will look like one of the following three assertions: (1) H_a: $\theta > \theta_0$ (in which case the implicit null hypothesis is $\theta \leq \theta_0$), (2) H_a: $\theta < \theta_0$ (so the implicit null hypothesis states that $\theta \geq \theta_0$), or (3) H_a: $\theta \neq \theta_0$. For example, let σ denote the standard deviation of the distribution of inside diameters (inches) for a certain type of metal sleeve. If the decision was made to use the sleeve unless sample evidence conclusively demonstrated that $\sigma > .001$, the appropriate hypotheses would be H_0: $\sigma = .001$ versus H_a: $\sigma > .001$. The number θ_0 that appears in both H_0 and H_a (separates the alternative from the null) is called the **null value.**

Test Procedures

A test procedure is a rule, based on sample data, for deciding whether to reject H_0. A test of H_0: $p = .10$ versus H_a: $p < .10$ in the circuit board problem might be based on examining a random sample of $n = 200$ boards. Let X denote the number of defective boards in the sample, a binomial random variable; x represents the observed value of X. If H_0 is true, $E(X) = np = 200(.10) = 20$, whereas we can expect

fewer than 20 defective boards if H_a is true. A value x just a bit below 20 does not strongly contradict H_0, so it is reasonable to reject H_0 only if x is substantially less than 20. One such test procedure is to reject H_0 if $x \leq 15$ and not reject H_0 otherwise. This procedure has two constituents: (1) a *test statistic* or function of the sample data used to make a decision and (2) a *rejection region* consisting of those x values for which H_0 will be rejected in favor of H_a. For the rule just suggested, the rejection region consists of $x = 0, 1, 2, \ldots,$ and 15. H_0 will not be rejected if $x = 16, 17, \ldots, 199,$ or 200.

A test procedure is specified by the following:

1. A **test statistic,** a function of the sample data on which the decision (reject H_0 or do not reject H_0) is to be based
2. A **rejection region,** the set of all test statistic values for which H_0 will be rejected

The null hypothesis will then be rejected if and only if the observed or computed test statistic value falls in the rejection region.

As another example, suppose a cigarette manufacturer claims that the average nicotine content μ of brand B cigarettes is (at most) 1.5 mg. It would be unwise to reject the manufacturer's claim without strong contradictory evidence, so an appropriate problem formulation is to test H_0: $\mu = 1.5$ versus H_a: $\mu > 1.5$. Consider a decision rule based on analyzing a random sample of 32 cigarettes. Let $\overline{X}$ denote the sample average nicotine content. If H_0 is true, $E(\overline{X}) = \mu = 1.5$, whereas if H_0 is false, we expect $\overline{X}$ to exceed 1.5. Strong evidence against H_0 is provided by a value $\overline{x}$ that considerably exceeds 1.5. Thus we might use $\overline{X}$ as a test statistic along with the rejection region $\overline{x} \geq 1.6$.

In both the circuit board and nicotine examples, the choice of test statistic and form of the rejection region make sense intuitively. However, the choice of cutoff value used to specify the rejection region is somewhat arbitrary. Instead of rejecting H_0: $p = .10$ in favor of H_a: $p < .10$ when $x \leq 15$, we could use the rejection region $x \leq 14$. For this region, H_0 would not be rejected if 15 defective boards are observed, whereas this occurrence would lead to rejection of H_0 if the initially suggested region is employed. Similarly, the rejection region $\overline{x} \geq 1.55$ might be used in the nicotine problem in place of the region $\overline{x} \geq 1.60$.

Errors in Hypothesis Testing

The basis for choosing a particular rejection region lies in consideration of the errors that one might be faced with in drawing a conclusion. Consider the rejection region $x \leq 15$ in the circuit board problem. Even when H_0: $p = .10$ is true, it might happen that an unusual sample results in $x = 13$, so that H_0 is erroneously rejected. On the other hand, even when H_a: $p < .10$ is true, an unusual sample might yield $x = 20$, in which case H_0 would not be rejected, again an incorrect conclusion. Thus it is possible that H_0 may be rejected when it is true or that H_0 may not be rejected when it is false. These possible errors are not consequences of a foolishly chosen rejection region. Either one of these two errors might result when the region $x \leq 14$ is employed, or indeed when any other region is used.

DEFINITION

A **type I error** consists of rejecting the null hypothesis H_0 when it is true.
A **type II error** involves not rejecting H_0 when H_0 is false.

In the nicotine problem, a type I error consists of rejecting the manufacturer's claim that $\mu = 1.5$ when it is actually true. If the rejection region $\bar{x} \geq 1.6$ is employed, it might happen that $\bar{x} = 1.63$ even when $\mu = 1.5$, resulting in a type I error. Alternatively, it may be that H_0 is false and yet $\bar{x} = 1.52$ is observed, leading to H_0 not being rejected (a type II error).

In the best of all possible worlds, test procedures for which neither type of error is possible could be developed. However, this ideal can be achieved only by basing a decision on an examination of the entire population. The difficulty with using a procedure based on sample data is that because of sampling variability, an unrepresentative sample may result. Even though $E(\bar{X}) = \mu$, the observed value $\bar{x}$ may differ substantially from μ (at least if n is small). Thus when $\mu = 1.5$ in the nicotine situation, $\bar{x}$ may be much larger than 1.5, resulting in erroneous rejection of H_0. Alternatively, it may be that $\mu = 1.6$ yet an $\bar{x}$ much smaller than this is observed, leading to a type II error.

Instead of demanding error-free procedures, we must look for procedures for which either type of error is unlikely to occur. That is, a good procedure is one for which the probability of making either type of error is small. The choice of a particular rejection region cutoff value fixes the probabilities of type I and type II errors. These error probabilities are traditionally denoted by α and β, respectively. Because H_0 specifies a unique value of the parameter, there is a single value of α. However, there is a different value of β for each value of the parameter consistent with H_a.

Example 8.1 A certain type of automobile is known to sustain no visible damage 25% of the time in 10-mph crash tests. A modified bumper design has been proposed in an effort to increase this percentage. Let p denote the proportion of all 10-mph crashes with this new bumper that result in no visible damage. The hypotheses to be tested are H_0: $p = .25$ (no improvement) versus H_a: $p > .25$. The test will be based on an experiment involving $n = 20$ independent crashes with prototypes of the new design. Intuitively, H_0 should be rejected if a substantial number of the crashes show no damage. Consider the following test procedure:

Test statistic: X = the number of crashes with no visible damage

Rejection region: $R_8 = \{8, 9, 10, \ldots, 19, 20\}$; that is, reject H_0 if $x \geq 8$, where x is the observed value of the test statistic.

This rejection region is called *upper-tailed* because it consists only of large values of the test statistic.

When H_0 is true, X has a binomial probability distribution with $n = 20$ and $p = .25$. Then

$$\begin{aligned}\alpha &= P(\text{type I error}) = P(H_0 \text{ is rejected when it is true}) \\ &= P(X \geq 8 \text{ when } X \sim \text{Bin}(20, .25)) = 1 - B(7; 20, .25) \\ &= 1 - .898 = .102\end{aligned}$$

That is, when H_0 is actually true, roughly 10% of all experiments consisting of 20 crashes would result in H_0 being incorrectly rejected (a type I error).

In contrast to α, there is not a single β. Instead, there is a different β for each different p that exceeds .25. Thus there is a value of β for $p = .3$ (in which case $X \sim \text{Bin}(20, .3)$), another value of β for $p = .5$, and so on. For example,

$$\begin{aligned}\beta(.3) &= P(\text{type II error when } p = .3)\\ &= P(H_0 \text{ is not rejected when it is false because } p = .3)\\ &= P(X \le 7 \text{ when } X \sim \text{Bin}(20, .3)) = B(7; 20, .3) = .772\end{aligned}$$

When p is actually .3 rather than .25 (a "small" departure from H_0), roughly 77% of all experiments of this type would result in H_0 being incorrectly not rejected!

The accompanying table displays β for selected values of p (each calculated for the rejection region R_8). Clearly, β decreases as the value of p moves farther to the right of the null value .25. Intuitively, the greater the departure from H_0, the less likely it is that such a departure will not be detected.

p	.3	.4	.5	.6	.7	.8
$\beta(p)$	.772	.416	.132	.021	.001	.000

The proposed test procedure is still reasonable for testing the more realistic null hypothesis that $p \le .25$. In this case, there is no longer a single α, but instead there is an α for each p that is at most .25: $\alpha(.25)$, $\alpha(.23)$, $\alpha(.20)$, $\alpha(.15)$, and so on. It is easily verified, though, that $\alpha(p) < \alpha(.25) = .102$ if $p < .25$. That is, the largest value of α occurs for the boundary value .25 between H_0 and H_a. Thus if α is small for the simplified null hypothesis, it will also be as small as or smaller for the more realistic H_0. ■

Example 8.2 The drying time of a certain type of paint under specified test conditions is known to be normally distributed with mean value 75 min and standard deviation 9 min. Chemists have proposed a new additive designed to decrease average drying time. It is believed that drying times with this additive will remain normally distributed with $\sigma = 9$. Because of the expense associated with the additive, evidence should strongly suggest an improvement in average drying time before such a conclusion is adopted. Let μ denote the true average drying time when the additive is used. The appropriate hypotheses are H_0: $\mu = 75$ versus H_a: $\mu < 75$. Only if H_0 can be rejected will the additive be declared successful and used.

Experimental data is to consist of drying times from $n = 25$ test specimens. Let $X_1, \ldots, X_{25}$ denote the 25 drying times—a random sample of size 25 from a normal distribution with mean value μ and standard deviation $\sigma = 9$. The sample mean drying time $\overline{X}$ then has a normal distribution with expected value $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 9/\sqrt{25} = 1.80$. When H_0 is true, $\mu_{\overline{X}} = 75$, so an $\overline{x}$ value somewhat less than 75 would not strongly contradict H_0. A reasonable rejection region has the form $\overline{X} \le c$, where the cutoff value c is suitably chosen. Consider the choice $c = 70.8$, so that the test procedure consists of test statistic $\overline{X}$ and rejection region $\overline{x} \le 70.8$. Because the rejection region consists only of small values of the test statistic, the test is said to be *lower-tailed.* Calculation of α and β now involves a routine standardization of $\overline{X}$ followed by reference to the standard normal probabilities of Appendix Table A.3:

$$\begin{aligned}\alpha &= P(\text{type I error}) = P(H_0 \text{ is rejected when it is true})\\ &= P(\overline{X} \le 70.8 \text{ when } \overline{X} \sim \text{normal with } \mu_{\overline{X}} = 75, \sigma_{\overline{X}} = 1.8)\\ &= \Phi\left(\frac{70.8 - 75}{1.8}\right) = \Phi(-2.33) = .01\end{aligned}$$

$$\begin{aligned}\beta(72) &= P(\text{type II error when } \mu = 72)\\ &= P(H_0 \text{ is not rejected when it is false because } \mu = 72)\\ &= P(\bar{X} > 70.8 \text{ when } \bar{X} \sim \text{normal with } \mu_{\bar{X}} = 72 \text{ and } \sigma_{\bar{X}} = 1.8)\\ &= 1 - \Phi\left(\frac{70.8 - 72}{1.8}\right) = 1 - \Phi(-.67) = 1 - .2514 = .7486\end{aligned}$$

$$\beta(70) = 1 - \Phi\left(\frac{70.8 - 70}{1.8}\right) = .3300 \qquad \beta(67) = .0174$$

For the specified test procedure, only 1% of all experiments carried out as described will result in H_0 being rejected when it is actually true. However, the chance of a type II error is very large when $\mu = 72$ (only a small departure from H_0), somewhat less when $\mu = 70$, and quite small when $\mu = 67$ (a very substantial departure from H_0). These error probabilities are illustrated in Figure 8.1. Notice that α is computed using the probability distribution of the test statistic when H_0 is true, whereas determination of β requires knowing the test statistic's distribution when H_0 is false.

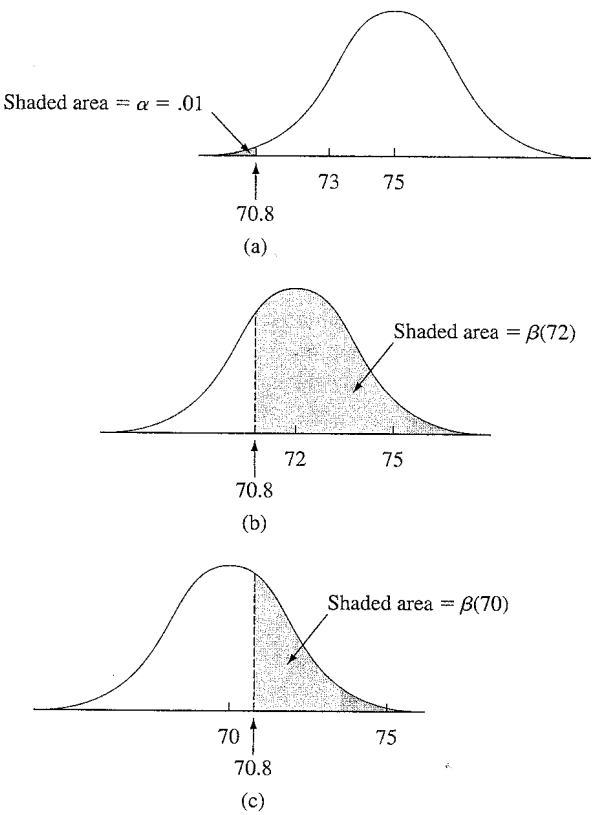

Figure 8.1 α and β illustrated for Example 8.2: (a) the distribution of $\bar{X}$ when $\mu = 75$ (H_0 true); (b) the distribution of $\bar{X}$ when $\mu = 72$ (H_0 false); (c) the distribution of $\bar{X}$ when $\mu = 70$ (H_0 false)

As in Example 8.1, if the more realistic null hypothesis $\mu \geq 75$ is considered, there is an α for each parameter value for which H_0 is true: $\alpha(75)$, $\alpha(75.8)$, $\alpha(76.5)$, and so on. It is easily verified, though, that $\alpha(75)$ is the largest of all these type I error probabilities. Focusing on the boundary value amounts to working explicitly with the "worst case." ■

The specification of a cutoff value for the rejection region in the examples just considered was somewhat arbitrary. Use of $R_8 = \{8, 9, \ldots, 20\}$ in Example 8.1 resulted in $\alpha = .102$, $\beta(.3) = .772$, and $\beta(.5) = .132$. Many would think these error probabilities intolerably large. Perhaps they can be decreased by changing the cutoff value.

Example 8.3 (Example 8.1 continued) Let us use the same experiment and test statistic X as previously described in the automobile bumper problem but now consider the rejection region $R_9 = \{9, 10, \ldots, 20\}$. Since X still has a binomial distribution with parameters $n = 20$ and p,

$$\begin{aligned}\alpha &= P(H_0 \text{ is rejected when } p = .25)\\ &= P(X \geq 9 \text{ when } X \sim \text{Bin}(20, .25)) = 1 - B(8; 20, .25) = .041\end{aligned}$$

The type I error probability has been decreased by using the new rejection region. However, a price has been paid for this decrease:

$$\begin{aligned}\beta(.3) &= P(H_0 \text{ is not rejected when } p = .3)\\ &= P(X \leq 8 \text{ when } X \sim \text{Bin}(20, .3)) = B(8; 20, .3) = .887\\ \beta(.5) &= B(8; 20, .5) = .252\end{aligned}$$

Both these βs are larger than the corresponding error probabilities .772 and .132 for the region R_8. In retrospect, this is not surprising; α is computed by summing over probabilities of test statistic values *in the rejection region,* whereas β is the probability that X falls *in the complement* of the rejection region. Making the rejection region smaller must therefore decrease α while increasing β for any fixed alternative value of the parameter. ■

Example 8.4 (Example 8.2 continued) The use of cutoff value $c = 70.8$ in the paint-drying example resulted in a very small value of α (.01) but rather large βs. Consider the same experiment and test statistic $\bar{X}$ with the new rejection region $\bar{x} \leq 72$. Because $\bar{X}$ is still normally distributed with mean value $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = 1.8$,

$$\begin{aligned}\alpha &= P(H_0 \text{ is rejected when it is true})\\ &= P(\bar{X} \leq 72 \text{ when } \bar{X} \sim N(75, 1.8^2))\\ &= \Phi\left(\frac{72 - 75}{1.8}\right) = \Phi(-1.67) = .0475 \approx .05\end{aligned}$$

$$\begin{aligned}\beta(72) &= P(H_0 \text{ is not rejected when } \mu = 72)\\ &= P(\bar{X} > 72 \text{ when } \bar{X} \text{ is a normal rv with mean 72 and standard deviation 1.8})\\ &= 1 - \Phi\left(\frac{72 - 72}{1.8}\right) = 1 - \Phi(0) = .5\end{aligned}$$

$$\beta(70) = 1 - \Phi\left(\frac{72 - 70}{1.8}\right) = .1335 \qquad \beta(67) = .0027$$

The change in cutoff value has made the rejection region larger (it includes more $\bar{x}$ values), resulting in a decrease in β for each fixed μ less than 75. However, α for this new region has increased from the previous value .01 to approximately .05. If a type I error probability this large can be tolerated, though, the second region ($c = 72$) is preferable to the first ($c = 70.8$) because of the smaller βs. ■

The results of these examples can be generalized in the following manner.

PROPOSITION

Suppose an experiment and a sample size are fixed and a test statistic is chosen. Then decreasing the size of the rejection region to obtain a smaller value of α results in a larger value of β for any particular parameter value consistent with H_a.

This proposition says that once the test statistic and n are fixed, there is no rejection region that will simultaneously make both α and all βs small. A region must be chosen to effect a compromise between α and β.

Because of the suggested guidelines for specifying H_0 and H_a, a type I error is usually more serious than a type II error (this can always be achieved by proper choice of the hypotheses). The approach adhered to by most statistical practitioners is then to specify the largest value of α that can be tolerated and find a rejection region having that value of α rather than anything smaller. This makes β as small as possible subject to the bound on α. The resulting value of α is often referred to as the **significance level** of the test. Traditional levels of significance are .10, .05, and .01, though the level in any particular problem will depend on the seriousness of a type I error—the more serious this error, the smaller should be the significance level. The corresponding test procedure is called a **level α test** (e.g., a level .05 test or a level .01 test). A test with significance level α is one for which the type I error probability is controlled at the specified level.

Example 8.5 Consider the situation mentioned previously in which μ was the true average nicotine content of brand B cigarettes. The objective is to test H_0: $\mu = 1.5$ versus H_a: $\mu > 1.5$ based on a random sample $X_1, X_2, \ldots, X_{32}$ of nicotine contents. Suppose the distribution of nicotine content is known to be normal with $\sigma = .20$. Then $\overline{X}$ is normally distributed with mean value $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = .20/\sqrt{32} = .0354$.

Rather than use $\overline{X}$ itself as the test statistic, let's standardize $\overline{X}$ assuming that H_0 is true.

$$\text{Test statistic:} \quad Z = \frac{\overline{X} - 1.5}{\sigma/\sqrt{n}} = \frac{\overline{X} - 1.5}{.0354}$$

Z expresses the distance between $\overline{X}$ and its expected value when H_0 is true as some number of standard deviations. For example, $z = 3$ results from an $\overline{x}$ that is 3 standard deviations larger than we would have expected it to be were H_0 true.

Rejecting H_0 when $\overline{x}$ "considerably" exceeds 1.5 is equivalent to rejecting H_0 when z "considerably" exceeds 0. That is, the form of the rejection region is $z \geq c$. Let's now determine c so that $\alpha = .05$. When H_0 is true, Z has a standard normal distribution. Thus

$$\begin{aligned} \alpha &= P(\text{type I error}) = P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true}) \\ &= P(Z \geq c \text{ when } Z \sim N(0, 1)) \end{aligned}$$

The value c must capture upper-tail area .05 under the z curve. Either from Section 4.3 or directly from Appendix Table A.3, $c = z_{.05} = 1.645$.

Notice that $z \geq 1.645$ is equivalent to $\overline{x} - 1.5 \geq (.0354)(1.645)$, that is, $\overline{x} \geq 1.56$. Then β is the probability that $\overline{X} < 1.56$ and can be calculated for any μ greater than 1.5. ■

a. Using this data, test at level .01 the null hypothesis that the company's premise is correct against the alternative that it is not correct.
b. What is the probability that when the test of part (a) is used, the company's premise will be judged correct when in fact 10% of all current customers qualify?

42. Each of a group of 20 intermediate tennis players is given two rackets, one having nylon strings and the other synthetic gut strings. After several weeks of playing with the two rackets, each player will be asked to state a preference for one of the two types of strings. Let p denote the proportion of all such players who would prefer gut to nylon, and let X be the number of players in the sample who prefer gut. Because gut strings are more expensive, consider the null hypothesis that at most 50% of all such players prefer gut. We simplify this to H_0: $p = .5$, planning to reject H_0 only if sample evidence strongly favors gut strings.
a. Which of the rejection regions $\{15, 16, 17, 18, 19, 20\}$, $\{0, 1, 2, 3, 4, 5\}$, or $\{0, 1, 2, 3, 17, 18, 19, 20\}$ is most appropriate, and why are the other two not appropriate?
b. What is the probability of a type I error for the chosen region of part (a)? Does the region specify a level .05 test? Is it the best level .05 test?
c. If 60% of all enthusiasts prefer gut, calculate the probability of a type II error using the appropriate region from part (a). Repeat if 80% of all enthusiasts prefer gut.
d. If 13 out of the 20 players prefer gut, should H_0 be rejected using a significance level of .10?

43. A manufacturer of plumbing fixtures has developed a new type of washerless faucet. Let $p = P$(a randomly selected faucet of this type will develop a leak within 2 years under normal use). The manufacturer has decided to proceed with production unless it can be determined that p is too large; the borderline acceptable value of p is specified as .10. The manufacturer decides to subject n of these faucets to accelerated testing (approximating 2 years of normal use). With X = the number among the n faucets that leak before the test concludes, production will commence unless the observed X is too large. It is decided that if $p = .10$, the probability of not proceeding should be at most .10, whereas if $p = .30$ the probability of proceeding should be at most .10. Can $n = 10$ be used? $n = 20$? $n = 25$? What is the appropriate rejection region for the chosen n, and what are the actual error probabilities when this region is used?

44. Scientists think that robots will play a crucial role in factories in the next several decades. Suppose that in an experiment to determine whether the use of robots to weave computer cables is feasible, a robot was used to assemble 500 cables. The cables were examined and there were 15 defectives. If human assemblers have a defect rate of .035 (3.5%), does this data support the hypothesis that the proportion of defectives is lower for robots than humans? Use a .01 significance level.

8.4 P-Values

One way to report the result of a hypothesis-testing analysis is to simply say whether the null hypothesis was rejected at a specified level of significance. Thus an investigator might state that H_0 was rejected at level of significance .05 or that use of a level .01 test resulted in not rejecting H_0. This type of statement is somewhat inadequate because it says nothing about whether the conclusion was a very close call or quite clearcut. A related difficulty is that such a report imposes the specified significance level on other decision makers. In many decision situations, individuals may have different views concerning the consequences of a type I or type II error. Each individual would then want to select his or her own significance level—some selecting $\alpha = .05$, others .01, and so on—and reach a conclusion accordingly. This could result in some individuals rejecting H_0 while others conclude that the data does not show a strong enough contradiction of H_0 to justify its rejection.

Example 8.14 The true average time to initial relief of pain for a best-selling pain reliever is known to be 10 min. Let μ denote the true average time to relief for a company's newly developed reliever. The company wishes to produce and market this reliever only if it provides quicker relief than the best-seller, so wishes to test H_0: $\mu = 10$ versus H_a: $\mu < 10$. Only if experimental evidence leads to rejection of H_0 will the new reliever be introduced. After weighing the relative seriousness of each type of error, a single level of significance must be agreed on and a decision—to reject H_0 and introduce the reliever or not to do so—made at that level.

Suppose the new reliever has been introduced. The company supports its claim of quicker relief by stating that, based on an analysis of experimental data, H_0: $\mu = 10$

was rejected in favor of H_a: $\mu < 10$ using level of significance $\alpha = .10$. Any individuals contemplating a switch to this new reliever would naturally want to reach their own conclusions concerning the validity of the claim. Individuals who are satisfied with the best-seller would view a type I error (concluding that the new product provides quicker relief when it actually does not) as serious so might wish to use $\alpha = .05, .01$, or even smaller levels. Unfortunately, the nature of the company's statement prevents an individual decision maker from reaching a conclusion at such a level. The company has imposed its own choice of significance level on others. The report could have been done in a manner that allowed each individual flexibility in drawing a conclusion at a personally selected α. ■

A *P-value* conveys much information about the strength of evidence against H_0 and allows an individual decision maker to draw a conclusion at any specified level α. Before we give a general definition, consider how the conclusion in a hypothesis-testing problem depends on the selected level α.

Example 8.15 The nicotine content problem discussed in Example 8.5 involved testing H_0: $\mu = 1.5$ versus H_a: $\mu > 1.5$. Because of the inequality in H_a, the rejection region is upper-tailed, with H_0 rejected if $z \geq z_\alpha$. Suppose $z = 2.10$. The accompanying table displays the rejection region for each of four different αs along with the resulting conclusion.

Level of Significance α	Rejection Region	Conclusion
.05	$z \geq 1.645$	Reject H_0
.025	$z \geq 1.96$	Reject H_0
.01	$z \geq 2.33$	Do not reject H_0
.005	$z \geq 2.58$	Do not reject H_0

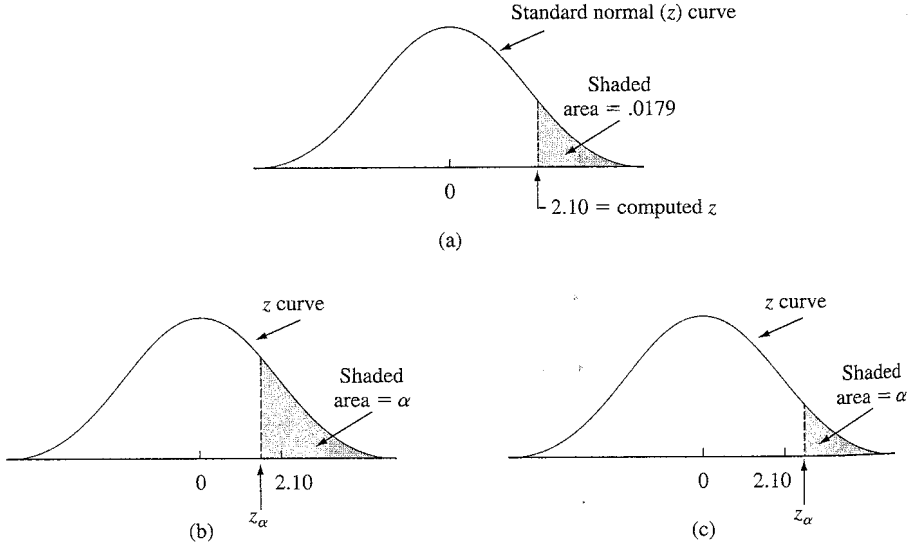

Figure 8.5 Relationship between α and tail area captured by computed z: (a) tail area captured by computed z; (b) when $\alpha > .0179$, $z_\alpha < 2.10$ and H_0 is rejected; (c) when $\alpha < .0179$, $z_\alpha > 2.10$ and H_0 is not rejected

For α relatively large, the z critical value z_α is not very far out in the upper tail; 2.10 exceeds the critical value, and so H_0 is rejected. However, as α decreases, the critical value increases. For small α, the z critical value is large, 2.10 is less than z_α, and H_0 is not rejected.

Recall that for an upper-tailed z test, α is just the area under the z curve to the right of the critical value z_α. That is, once α is specified, the critical value is chosen to capture upper-tail area α. Appendix Table A.3 shows that the area to the right of 2.10 is .0179. Using an α larger than .0179 corresponds to $z_\alpha < 2.10$. An α less than .0179 necessitates using a z critical value that exceeds 2.10. The decision at a particular level α thus depends on how the selected α compares to the tail area captured by the computed z. This is illustrated in Figure 8.5. Notice in particular that .0179, the captured tail area, is the smallest level α at which H_0 would be rejected, because using any smaller α results in a z critical value that exceeds 2.10, so that 2.10 is not in the rejection region. ■

In general, suppose the probability distribution of a test statistic when H_0 is true has been determined. Then, for specified α, the rejection region is determined by finding a critical value or values that capture tail area α (upper-, lower-, or two-tailed, whichever is appropriate) under the probability distribution curve. The smallest α for which H_0 would be rejected is the tail area captured by the computed value of the test statistic. This smallest α is the P-value.

DEFINITION

The ***P*-value** (or *observed significance level*) is the smallest level of significance at which H_0 would be rejected when a specified test procedure is used on a given data set. Once the P-value has been determined, the conclusion at any particular level α results from comparing the P-value to α:

1. P-value $\leq \alpha \Rightarrow$ reject H_0 at level α.
2. P-value $> \alpha \Rightarrow$ do not reject H_0 at level α.

It is customary to call the data *significant* when H_0 is rejected and *not significant* otherwise. The P-value is then the smallest level at which the data is significant. An easy way to visualize the comparison of the P-value with the chosen α is to draw a picture like that of Figure 8.6. The calculation of the P-value depends on whether the test is upper-, lower-, or two-tailed. However, once it has been calculated, the comparison with α does not depend on which type of test was used.

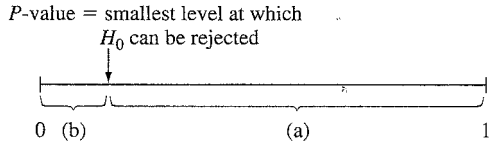

Figure 8.6 Comparing α and the P-value: (a) reject H_0 when α lies here; (b) do not reject H_0 when α lies here

Example 8.16 (Example 8.14 continued) Suppose that when data from an experiment involving the new pain reliever was analyzed, the P-value for testing H_0: $\mu = 10$ versus H_a: $\mu < 10$ was calculated as .0384. Since $\alpha = .05$ is larger than the P-value (.05 lies in the interval (a) of Figure 8.6),

H_0 would be rejected by anyone carrying out the test at level .05. However, at level .01, H_0 would not be rejected because .01 is smaller than the smallest level (.0384) at which H_0 can be rejected. ■

The most widely used statistical computer packages automatically include a P-value when a hypothesis-testing analysis is performed. A conclusion can then be drawn directly from the output, without reference to a table of critical values.

A useful alternative definition equivalent to the one just given is as follows:

DEFINITION

> The ***P*-value** is the probability, calculated assuming H_0 is true, of obtaining a test statistic value at least as contradictory to H_0 as the value that actually resulted. The smaller the P-value, the more contradictory is the data to H_0.

Thus if $z = 2.10$ for an upper-tailed z test, P-value $= P(Z \geq 2.10$ when H_0 is true$) = 1 - \Phi(2.10) = .0179$, as before. Beware: The P-value is not the probability that H_0 is true, nor is it an error probability!

P-Values for *z* Tests

The P-value for a z test (one based on a test statistic whose distribution when H_0 is true is at least approximately standard normal) is easily determined from the information in Appendix Table A.3. Consider an upper-tailed test and let z denote the computed value of the test statistic Z. The null hypothesis is rejected if $z \geq z_\alpha$, and the P-value is the smallest α for which this is the case. Since z_α increases as α decreases, the P-value is the value of α for which $z = z_\alpha$. That is, the P-value is just the area captured by the computed value z in the upper tail of the standard normal curve. The corresponding cumulative area is $\Phi(z)$, so in this case P-value $= 1 - \Phi(z)$.

An analogous argument for a lower-tailed test shows that the P-value is the area captured by the computed value z in the lower tail of the standard normal curve. More care must be exercised in the case of a two-tailed test. Suppose first that z is positive. Then the P-value is the value of α satisfying $z = z_{\alpha/2}$ (i.e., computed z = upper-tail critical value). This says that the area captured in the upper tail is half the P-value, so that P-value $= 2[1 - \Phi(z)]$. If z is negative, the P-value is the α for which $z = -z_{\alpha/2}$, or, equivalently, $-z = z_{\alpha/2}$, so P-value $= 2[1 - \Phi(-z)]$. Since $-z = |z|$ when z is negative, P-value $= 2[1 - \Phi(|z|)]$ for either positive or negative z.

$$P\text{-value: } P = \begin{cases} 1 - \Phi(z) & \text{for an upper-tailed test} \\ \Phi(z) & \text{for a lower-tailed test} \\ 2[1 - \Phi(|z|)] & \text{for a two-tailed test} \end{cases}$$

Each of these is the probability of getting a value at least as extreme as what was obtained (assuming H_0 true). The three cases are illustrated in Figure 8.7.

The next example illustrates the use of the P-value approach to hypothesis testing by means of a sequence of steps modified from our previously recommended sequence.

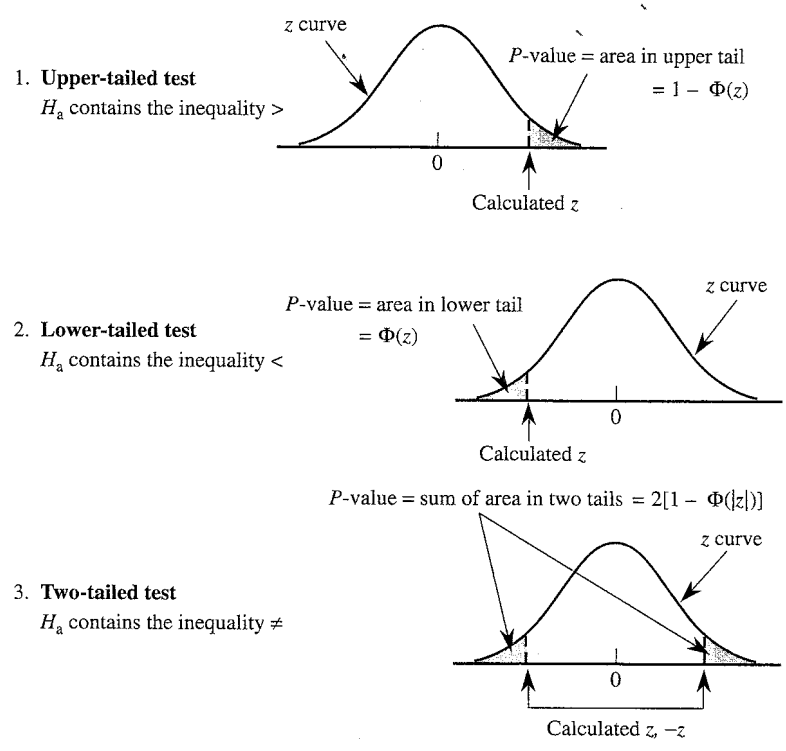

Figure 8.7 Determination of the *P*-value for a *z* test

Example 8.17 The target thickness for silicon wafers used in a certain type of integrated circuit is 245 μm. A sample of 50 wafers is obtained and the thickness of each one is determined, resulting in a sample mean thickness of 246.18 μm and a sample standard deviation of 3.60 μm. Does this data suggest that true average wafer thickness is something other than the target value?

1. Parameter of interest: μ = true average wafer thickness
2. Null hypothesis: H_0: $\mu = 245$
3. Alternative hypothesis: H_a: $\mu \neq 245$
4. Formula for test statistic value: $z = \dfrac{\bar{x} - 245}{s/\sqrt{n}}$
5. Calculation of test statistic value: $z = \dfrac{246.18 - 245}{3.60/\sqrt{50}} = 2.32$
6. Determination of *P*-value: Because the test is two-tailed,
$$P\text{-value} = 2(1 - \Phi(2.32)) = .0204$$
7. Conclusion: Using a significance level of .01, H_0 would not be rejected since .0204 > .01. At this significance level, there is insufficient evidence to conclude that true average thickness differs from the target value. ■

P-Values for *t* Tests

Just as the *P*-value for a *z* test is a *z* curve area, the *P*-value for a *t* test will be a *t* curve area. Figure 8.8 on the next page illustrates the three different cases. The number of df for the one-sample *t* test is $n - 1$.

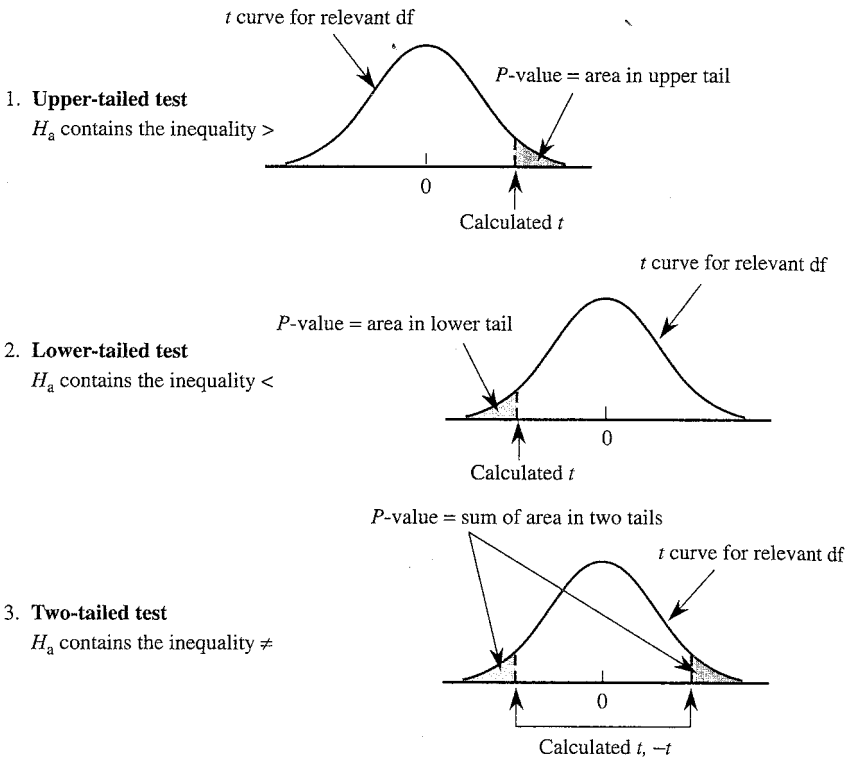

Figure 8.8 *P*-values for *t* tests

The table of t critical values used previously for confidence and prediction intervals doesn't contain enough information about any particular t distribution to allow for accurate determination of desired areas. So we have included another t table in Appendix Table A.8, one that contains a tabulation of upper-tail t curve areas. Each different column of the table is for a different number of df, and the rows are for calculated values of the test statistic t ranging from 0.0 to 4.0 in increments of .1. For example, the number .074 appears at the intersection of the 1.6 row and the 8 df column, so the area under the 8 df curve to the right of 1.6 (an upper-tail area) is .074. Because t curves are symmetric, .074 is also the area under the 8 df curve to the left of −1.6 (a lower-tail area).

Suppose, for example, that a test of H_0: $\mu = 100$ versus H_a: $\mu > 100$ is based on the 8 df t distribution. If the calculated value of the test statistic is $t = 1.6$, then the P-value for this upper-tailed test is .074. Because .074 exceeds .05, we would not be able to reject H_0 at a significance level of .05. If the alternative hypothesis is H_a: $\mu < 100$ and a test based on 20 df yields $t = -3.2$, then Appendix Table A.8 shows that the P-value is the captured lower-tail area .002. The null hypothesis can be rejected at either level .05 or .01. Consider testing H_0: $\mu_1 - \mu_2 = 0$ versus H_a: $\mu_1 - \mu_2 \neq 0$; the null hypothesis states that the means of the two populations are identical, whereas the alternative hypothesis states that they are different without specifying a direction of departure from H_0. If a t test is based on 20 df and $t = 3.2$, then the P-value for this two-tailed test is 2(.002) = .004. This would also be the P-value for $t = -3.2$. The tail area is doubled because values both larger than 3.2 and smaller than −3.2 are more contradictory to H_0 than what was calculated (values farther out in *either* tail of the t curve).

Example 8.18 In Example 8.9, we carried out a test of H_0: $\mu = 25$ versus H_a: $\mu > 25$ based on 4 df. The calculated value of t was 1.04. Looking to the 4 df column of Appendix Table A.8 and down to the 1.0 row, we see that the entry is .187, so P-value $\approx$ .187. This P-value is clearly larger than any reasonable significance level α (.01, .05, and even .10), so there is no reason to reject the null hypothesis. The MINITAB output included in Example 8.9 has P-value = .18. P-values from software packages will be more accurate than what results from Appendix Table A.8 since values of t in our table are accurate only to the tenths digit. ■

EXERCISES Section 8.4 (45–60)

45. For which of the given P-values would the null hypothesis be rejected when performing a level .05 test?
a. .001 **b.** .021 **c.** .078
d. .047 **e.** .148

46. Pairs of P-values and significance levels, α, are given. For each pair, state whether the observed P-value would lead to rejection of H_0 at the given significance level.
a. P-value = .084, α = .05
b. P-value = .003, α = .001
c. P-value = .498, α = .05
d. P-value = .084, α = .10
e. P-value = .039, α = .01
f. P-value = .218, α = .10

47. Let μ denote the mean reaction time to a certain stimulus. For a large-sample z test of H_0: $\mu = 5$ versus H_a: $\mu > 5$, find the P-value associated with each of the given values of the z test statistic.
a. 1.42 **b.** .90 **c.** 1.96 **d.** 2.48 **e.** −.11

48. Newly purchased tires of a certain type are supposed to be filled to a pressure of 30 lb/in^2. Let μ denote the true average pressure. Find the P-value associated with each given z statistic value for testing H_0: $\mu = 30$ versus H_a: $\mu \neq 30$.
a. 2.10 **b.** −1.75 **c.** −.55 **d.** 1.41 **e.** −5.3

49. Give as much information as you can about the P-value of a t test in each of the following situations:
a. Upper-tailed test, df = 8, t = 2.0
b. Lower-tailed test, df = 11, t = −2.4
c. Two-tailed test, df = 15, t = −1.6
d. Upper-tailed test, df = 19, t = −.4
e. Upper-tailed test, df = 5, t = 5.0
f. Two-tailed test, df = 40, t = −4.8

50. The paint used to make lines on roads must reflect enough light to be clearly visible at night. Let μ denote the true average reflectometer reading for a new type of paint under consideration. A test of H_0: $\mu = 20$ versus H_a: $\mu > 20$ will be based on a random sample of size n from a normal population distribution. What conclusion is appropriate in each of the following situations?
a. n = 15, t = 3.2, α = .05
b. n = 9, t = 1.8, α = .01
c. n = 24, t = −.2

51. Let μ denote true average serum receptor concentration for all pregnant women. The average for all women is known to be 5.63. The article "Serum Transferrin Receptor for the Detection of Iron Deficiency in Pregnancy" (*Amer. J. Clinical Nutr.*, 1991: 1077–1081) reports that P-value > .10 for a test of H_0: $\mu = 5.63$ versus H_a: $\mu \neq 5.63$ based on n = 176 pregnant women. Using a significance level of .01, what would you conclude?

52. The article "Analysis of Reserve and Regular Bottlings: Why Pay for a Difference Only the Critics Claim to Notice?" (*Chance*, Summer 2005, pp. 9–15) reported on an experiment to investigate whether wine tasters could distinguish between more expensive reserve wines and their regular counterparts. Wine was presented to tasters in four containers labeled A, B, C, and D, with two of these containing the reserve wine and the other two the regular wine. Each taster randomly selected three of the containers, tasted the selected wines, and indicated which of the three he/she believed was different from the other two. Of the n = 855 tasting trials, 346 resulted in correct distinctions (either the one reserve that differed from the two regular wines or the one regular wine that differed from the two reserves). Does this provide compelling evidence for concluding that tasters of this type have some ability to distinguish between reserve and regular wines? State and test the relevant hypotheses using the P-value approach. Are you particularly impressed with the ability of tasters to distinguish between the two types of wine?

53. An aspirin manufacturer fills bottles by weight rather than by count. Since each bottle should contain 100 tablets, the average weight per tablet should be 5 grains. Each of 100 tablets taken from a very large lot is weighed, resulting in a sample average weight per tablet of 4.87 grains and a sample standard deviation of .35 grain. Does this information provide strong evidence for concluding that the company is not filling its bottles as advertised? Test the appropriate hypotheses using α = .01 by first computing the P-value and then comparing it to the specified significance level.

54. Because of variability in the manufacturing process, the actual yielding point of a sample of mild steel subjected to increasing stress will usually differ from the theoretical yielding point. Let p denote the true proportion of samples that yield before their theoretical yielding point. If on the